Lecture #18

- Last time, we analyzed the issue of independence of two binary r.v. (X, Y). We did this through a contingency table. This is easily generalized to a non-binary but still finite r.v. by simply increasing the number of rows and columns in the table.

- Suppose X takes values in $\{1, 2, \dots, m\}$ and Y takes values in $\{1, 2, \dots, n\}$ (we shall reserve N for total # samples in what follows). Suppose we make N observations jointing from (X, Y), call them $\{(x_i, y_i)\}_{i=1}^{N}$. Generalizing the binary case, let

$$N_{jk} := \#\text{ observations of the form } (j, k)$$

- Build a contingency table:

	$X=0$	$X=1$	$\dots$	$X=m$	
$Y=0$	N_{00}	N_{10}		N_{m0}	$N_{\cdot 0}$
$Y=1$	N_{01}	N_{11}		N_{m1}	$N_{\cdot 1}$
$\vdots$					
$Y=n$	N_{0n}	N_{1n}		N_{mn}	$N_{\cdot m}$
	$N_{0\cdot}$	$N_{1\cdot}$		$N_{m\cdot}$	N

Consider, as in the binary case, hypotheses $H_0: X \perp\!\!\!\perp Y$

$H_1: X \not\perp\!\!\!\perp Y$

- We can perform a similar test as in the binary case, albeit with more d.o.f. in the χ^2 distribution.

Theorem (χ^2 Test for Independence): Let $E_{jk} = \frac{(N_{j\cdot})(N_{\cdot k})}{N}$. Then under $H_0: X \perp\!\!\!\perp Y$, the test statistic $U = \sum_{j=1}^{m} \sum_{k=1}^{n} \frac{[N_{jk} - E_{jk}]^2}{E_{jk}}$

is distributed as a χ^2 r.v. with $[m-1]\cdot[n-1]$ degrees of freedom.

ex: Consider the random variables corresponding to political party affiliation and amount donated in most recent election

$X = \{D, G, L, R\}$

$Y = \{\{0\}, (1,50), (51,1000), (1000,\infty)\}$

	$\{0\}$	$(1,50)$	$(51,1000)$	$(1000,\infty)$	
D	78	41	20	9	148
G	4	2	0	0	6
L	17	4	3	1	25
R	51	61	22	6	140
	150	108	45	16	319

One can compute a corresponding "table of expectations"

	69.59	50.11	20.88	7.43
	2.82	2.03	0.85	0.30
	11.76	8.46	3.53	1.25
	65.83	47.40	19.75	7.02

Computing the test statistic $U = \sum_{j,k}^{4} \frac{[O_{jk} - E_{jk}]^2}{E_{jk}}$

$= 17.16$

So, what happens if we consider H_0: [Party Affiliation] $\perp\!\!\!\perp$ [Giving Amount]

H_1: [Party Affiliation] $\not\perp\!\!\!\perp$ [Giving Amount]

at the $\alpha = .05$ level? Well, the cut-off for $\alpha = .05$ at d.o.f. $= [4-1]\cdot[4-1]$
$= 9$

is 16.92. So, we barely reject H_0.

Remark: The way the donation amount bins were designed plays an important role in such tests, but they are somewhat arbitrary. Be careful!

- One way around this strange issue is to allow for X to be discrete and Y continuous.
- Indeed, let X take values in $\{1, \dots, m\}$ and let Y be continuous, taking values in $\mathbb{R}$.
- Let $F_j(y) := \mathbb{P}(Y \le y \mid X = j)$. It is not hard to see that $\{F_j\}_{j=1}^m$ characterize the independence of X and Y:

Theorem: When X, Y are as above, then $X \perp\!\!\!\perp Y \Leftrightarrow F_j = F_k \ \forall j,k$.

- Of course, we are still interested in the case where $\{F_j\}_{j=1}^m$ are not known exactly, but must be estimated from samples. We handle this with the Kolmogorov-Smirnov test

- To simplify matters, suppose $m=2$, so X is binary, taking values $\{1,2\}$. Let us consider the null distribution that X and Y are independent, i.e.

$$H_0 : F_1 = F_2$$
$$H_1 : F_1 \neq F_2$$

- Suppose we take N samples from the joint distribution (X, Y): $\{(x_i, y_i)\}_{i=1}^N$.
Let $N_1 :=$ # samples with $X=1$, similarly for N_2.

- Define the empirical cdfs for Y conditioned on X:

$$\hat{F}_1(y) := \frac{1}{N_1} \cdot \sum_{i=1}^{N} \mathbb{1}(y_i \le y) \cdot \mathbb{1}(x_i = 1)$$

$$\hat{F}_2(y) := \frac{1}{N_2} \cdot \sum_{i=1}^{N} \mathbb{1}(y_i \le y) \cdot \mathbb{1}(x_i = 2)$$

If $F_1(y) \equiv F_2(y) \Leftrightarrow F_1(y) - F_2(y) \equiv 0$, then we have independence.

- Let $D = \sup_y \left| \hat{F}_1(y) - \hat{F}_2(y) \right|$ be an empirical measure of non-independence.

- We want to know how D behaves under $H_0: X \perp\!\!\!\perp Y$. It can be done, but this is not easy.

Theorem (K-S Test): Define

$$H(t) := 1 - 2 \sum_{r=1}^{\infty} (-1)^{r-1} \cdot \exp\left(-2r^2 t^2\right)$$

Under $H_0: F_1 = F_2$, $\lim_{N \to \infty} \mathbb{P}\left(\sqrt{\frac{N_1 N_2}{N_1 + N_2}} \cdot D \leq t \right) = H(t)$

Remarks:

- $H(t)$ is related to the Jacobi Theta function.

- The normalization constant $\frac{\sqrt{\frac{N_2 N_1}{N_2 + N_2}}}{\sqrt{N}}$ "converges" as $N \to \infty$, since

$$\sqrt{\frac{N_1 N_2}{N_1 + N_2}} = \sqrt{\frac{(N - N_1) N_1}{N}}$$

$$\rightsquigarrow \sqrt{\frac{(N - pN)(pN)}{N}}$$

$$= \sqrt{N} \cdot \left[\sqrt{(1-p)p}\right]$$

- To use this test, one needs to be able to compute $H(t)$ for many t values... it can be done! $H(t)$ is well-studied.